Unifying Annotated Discourse Hierarchies to Create a Gold Standard

Marco Carbone, Ya'akov Gal, Stuart Shieber, and **Barbara Grosz**
Division of Engineering and Applied Sciences
Harvard University
33 Oxford St.
Cambridge, MA 02138
mcarbone,gal,shieber,grosz@eecs.harvard.edu

Abstract

Human annotation of discourse corpora typically results in segmentation hierarchies that vary in their degree of agreement. This paper presents several techniques for unifying multiple discourse annotations into a single hierarchy, deemed a "gold standard" — the segmentation that best captures the underlying linguistic structure of the discourse. It proposes and analyzes methods that consider the level of embeddedness of a segmentation as well as methods that do not. A corpus containing annotated hierarchical discourses, the Boston Directions Corpus, was used to evaluate the "goodness" of each technique, by comparing the similarity of the segmentation it derives to the original annotations in the corpus. Several metrics of similarity between hierarchical segmentations are computed: precision/recall of matching utterances, pairwise inter-reliability scores (κ), and non-crossing-brackets. A novel method for unification that minimizes conflicts among annotators outperforms methods that require consensus among a majority for the κ and precision metrics, while capturing much of the structure of the discourse. When high recall is preferred, methods requiring a majority are preferable to those that demand full consensus among annotators.

1 Introduction

The linguistic structure of a discourse is composed of utterances that exhibit meaningful hierarchical relationships (Grosz and Sidner, 1986). Automatic segmentation of discourse forms the basis for many applications, from information retrieval and text summarization to anaphora resolution (Hearst, 1997). These automatic methods, usually based on supervised machine learning techniques, require a manually annotated corpus of data for training. The creation of these corpora often involves multiple judges annotating the same discourses, so as to avoid bias from using a single judge's annotations as ground truth. Usually, for a particular discourse, these multiple annotations are unified into a single annotation, either manually by the annotators' discussions or automatically. However, annotation unification approaches have not been formally evaluated, and although manual unification might be the best approach, it can be time-consuming. Indeed, much of the work on automatic recognition of discourse structure has focused on linear, rather than hierarchical segmentation (Hearst, 1997; Hirschberg and Nakatani, 1996), because of the difficulties of obtaining consistent hierarchical annotations. In addition, those approaches that do handle hierarchical segmentation do not address automatic unification methods (Carlson et al., 2001; Marcu, 2000).

There are several reasons for the prevailing emphasis on linear annotation and the lack of work on automatic methods for unifying hierarchical discourse annotations. First, initial attempts to create annotated hierarchical corpora of discourse structure using naive annotators have met with difficulties. Rotondo (1984) reported that "hierarchical segmentation is impractical for naive subjects in discourses longer than 200 words." Passonneau and Litman (1993) conducted a pilot study in which subjects found it "difficult and time-consuming" to identify hierarchical relations in discourse. Other attempts have had more success using improved annotation tools and more precise instructions (Grosz and Hirschberg, 1992; Hirschberg and Nakatani, 1996). Second, hierarchical segmentation of discourse is subjective. While agreement among annotators regarding linear segmentation has been found to be higher than 80% (Hearst, 1997), with respect to hierarchical segmentation it has been observed to be

as low as 60% (Flammia and Zue, 1995). Moreover, the precise definition of "agreement" with respect to hierarchical segmentation is unclear, complicating evaluation. It is natural to consider two segments in separate annotations to agree if they both span precisely the same utterances and agree on the level of embeddedness. However, it is less clear how to handle segments that share the same utterances but differ with respect to the level of embeddedness.

In this paper, we show that despite these difficulties it is possible to automatically combine a set of multi-level discourse annotations together into a single gold standard, a segmentation that best captures the underlying linguistic structure of the discourse. We aspire to create corpora analogous to the Penn Treebank in which a unique parse tree exists for each sentence that is agreed upon by all to convey the "correct" parse of the sentence. However, whereas the Penn Treebank parses are determined through a time-consuming negotiation between labelers, we aim to derive gold standard segmentations automatically.

There are several potential benefits for having a unifying standard for discourse corpora. First, it can constitute a unique segmentation of the discourse that is deemed the nearest approximation of the true objective structure, assuming one exists. Second, it can be used as a single unified version with which to train and evaluate algorithms for automatic discourse segmentation. Third, it can be used as a preprocessing step for computational tasks that require discourse structure, such as anaphora resolution and summarization.

In this work, we describe and evaluate several approaches for unifying multiple hierarchical discourse segmentations into one gold standard. Some of our approaches measure the agreement between annotations by taking into account the level of embeddedness and others ignore the hierarchy. We also introduce a novel method, called the Conflict-Free Union, that minimizes the number of conflicts between annotations. For our experiments, we used the Boston Directions Corpus (BDC).[1]

Ideally, each technique would be evaluated with respect to a single unified segmentation of the BDC that was deemed "true" by annotators who are experts in discourse theory, but we did not have the resources to attempt this task. Instead, we measure each technique by comparing the average similarity between its gold standard and the original annotations used to create it. Our similarity metrics measure both hierarchical and linear segment agreement using precision/recall metrics, inter-reliability similarities among annotations using the (κ) metric, and percentage of non-crossing-brackets.

We found that there is no single approach that does well with respect to all of the similarity metrics. However, the Conflict-Free Union approach outperforms the other methods for the κ and precision metrics. Also, techniques that include majority agreements of annotators have better recall than techniques which demanded full consensus among annotators. We also uncovered some flaws in each technique; for example, we found that gold standards that include dense structure are over-penalized by some of the metrics.

2 Methods for Creating a Gold Standard

It is likely that there is no perfect way to find and evaluate a gold standard, and in some cases there may be multiple segmentations that are equally likely to serve as a gold standard. In the BDC corpus, unlike the Penn Treebank, there are multiple annotations for each discourse which were not manually combined into one gold standard annotation. In this paper, we explore automatic methods to create a gold standard for the BDC corpus. These methods could also be used on other corpora with non-unified annotations. Next, we present several automatic methods to combine multiple human-annotated discourse segmentations into one gold standard.

2.1 Flat vs. Hierarchical Approaches

Most previous work that has combined multiple annotations has used linear segmentations, i.e. discourse segmentations without hierarchies (Hirschberg and Nakatani, 1996). In general, the hierarchical nature of discourse structure has not been considered when computing labeler inter-reliability and in evaluations of agreement with automatic methods. Since computational discourse theory relies on the hierarchy of its segments, we will consider it in this paper. For each approach that follows, we consider both a "flat" version, which does not consider level of embeddedness, and a "full" approach, which does. We analyze the differences between the flat and full versions for each approach.

2.2 Segment Definition

A discourse is made up of a sequence of utterances, $u_1, u_2, u_3, ..., u_n$. In this paper, we define a *segment* as a triple $\langle i, j, l \rangle$, where u_i is the first utterance in the segment, u_j is the last utterance in the segment, and l is the segment's level of embeddedness.[2] We will sometimes refer to u_i and u_j as *boundary utterances*. Lastly, when we are not interested in level of embeddedness, we will sometimes refer to a segment as $\langle i, j \rangle$, without the l value.

2.3 The Consensus Approach

A conservative way to combine segmentations into a gold standard is the Consensus (CNS) (or raw agreement) ap-

[1] The Boston Directions Corpus was designed and collected by Barbara Grosz, Julia Hirschberg, and Christine H. Nakatani.

[2] The levels are numbered from top to bottom; hence, 1 is the level of the largest segment, 2 is the level below that, and so on.

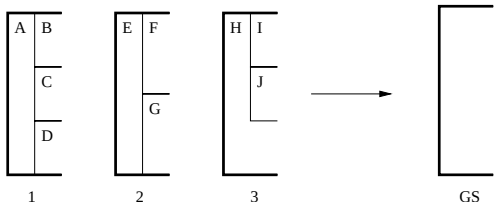

Figure 1: An example of the FullCNS approach. FlatCNS would create the same gold standard. The segments in the annotations that are marked in bold are those selected by the gold standard.

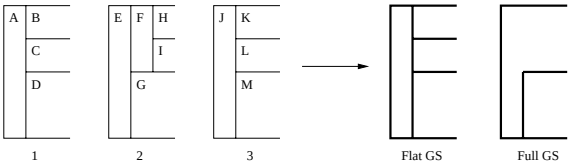

Figure 2: An example where FullCNS and FlatCNS create different gold standards..

proach (Hirschberg and Nakatani, 1996). CNS constructs a gold standard by including only those segments which have been included by every annotator. In the "full" version of CNS (FullCNS), the annotators need to agree upon the embedded level of the segment along with the segment boundaries. The "flat" version (FlatCNS) ignores hierarchy and considers only the segment boundaries when determining agreement.

Figure 1 shows an example of performing FullCNS on three annotations. In the figure, all three annotators agree on only the largest segment (segments A, E, and H). Hence, the gold standard includes only that single segment. FlatCNS gives the same gold standard in this example as there are no two segments with the same boundaries but with different levels of embeddedness. In Figure 2, we see an example where the gold standards created by FlatCNS and FullCNS differ. Aside from the largest segment, FullCNS contains only the segment representing the agreement of segments D, G, and M. FlatCNS includes that segment as well, in addition to two more from the agreement of segments B, H, and K and segments C, I, and L. FullCNS does not include those segments because the segments occur at differing levels of embeddedness.

2.4 Majority Consensus

A straightforward extension to the CNS approach is to relax the need for full agreement and include those segments on which a *majority* of the annotators agreed (Grosz and Hirschberg, 1992). Other thresholds of agreement could be used as well, but in this paper we

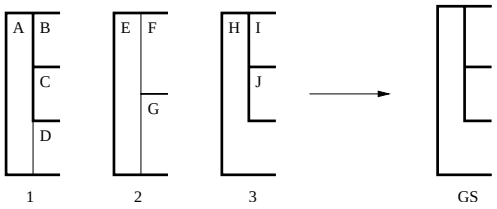

Figure 3: An example of the FullMCNS approach.

only consider it an agreement when more than 50% of annotators agree on a segment. We call this the Majority Consensus (MCNS) approach. As with CNS, we can have both a "full" version (FullMCNS) and a "flat" version (FlatMCNS), which do and do not consider the level of embeddedness, respectively.

Figure 3 shows an example of performing FullMCNS on the same three annotations we saw in Figure 1. Here, we again include the largest segment because it is agreed upon by all, but now we also include the two segments agreed upon by annotators 1 and 3 because two out of three annotators, a majority, have selected them. These two segments correspond to segments B and C for annotator 1 and segments I and J for annotator 3.

MCNS is less strict than CNS as it includes segments agreed upon by most annotators and does not require full agreement, but both methods are affected by a potential flaw. Note that in Figure 3, segment D could very well be in some notion of agreement with annotation 3, but MCNS does not capture this near-miss; D is left out of the gold standard. The next approach we discuss can handle this sort of situation.

2.5 Conflict-Free Union

The Conflict-Free Union (CFU) approach combines the annotations of all of the annotators and removes those segments that conflict with each other to get a conflict-free gold standard. There are usually multiple ways to construct a conflict-free gold standard. The CFU approach finds the one with the *fewest* segments removed.

Figure 4 shows the use of CFU on the three example annotations. Notice that the only segments not included in the gold standard are F and G, which conflict with B, C, D, I, and J. Resolving the conflicts here required removing two segments; the other way to resolve the conflict would have been to remove C and J, which would be equally as good. CFU captures as many conflict-free segments from the annotations as possible without discrimination. Even if only one annotator chose a segment, CFU would include it if it did not create more conflicts. Hence, it is likely that CFU could construct gold standards with too much structure. However, in our example it is better at capturing the similarity of structure between annotators 1 and 3. Due to its ability to capture structure,

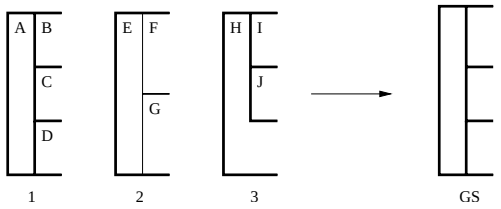

Figure 4: An example of the CFU approach.

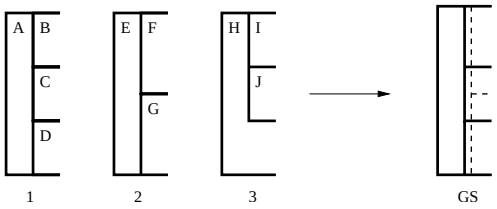

Figure 5: An example of the FullUNI approach.

we expected that CFU would perform better in recall than the previously mentioned approaches.

2.5.1 CFU Algorithm

The consensus and majority approaches are straightforward to compute, but CFU presents an optimization problem in which the greatest number of segments that can be combined without any internal conflicts must be found. Brute force methods, such as trying every possible set of segments and picking the largest conflict-free set, grow exponentially in the total number of segments contained in the annotations. We present a dynamic programming algorithm that computes the CFU in $O(n^3)$ time, where n is the number of utterances in the discourse.

First, we say that segment $\langle i, j \rangle$ *straddles* an utterance u_k when $i \leq k \leq j$. Let S_{ijk} represent the number of segments between utterances u_i and u_j, inclusive, that straddle u_k. That is, S_{ijk} represents the number of unique segments with the form $\langle x, y \rangle$, where $i \leq x \leq k \leq y \leq j$ and $x < y$. We use S_{ijk} to compute K_{ij}, the index representing the utterance u_k that, if considered a new boundary utterance, would minimize the number of conflicting segments within u_i and u_j, and T_{ij}, that minimum number of segments. Then we can solve the following recurrence equations:

$$T_{ii} = 0$$

$$T_{ij} = \min_{k=i}^{j}(T_{ik} + T_{(k+1)j} + S_{ijk}) \quad \text{where } j \geq i+1$$

$$K_{ij} = \arg\min_{k=i}^{j}(T_{ik} + T_{(k+1)j} + S_{ijk})$$

We can generate a binary tree with $\langle 1, N \rangle$ as the value of the root node, representing the segment over all utterances. The left child, then, has the value $\langle 1, K_{1N} \rangle$, and the right child has value $\langle K_{1N} + 1, N \rangle$. We compute the rest of the tree similarly, with $\langle 1, K_{1K_{1N}} \rangle$ as the left child of the left child, and so on. For each segment included by an annotator, we include it in the gold standard if it is represented by a segment in the constructed tree. Note that we only store the boundary utterances in the tree, so the gold standard we construct will not include level of embeddedness.

2.6 Union

The methods for finding a gold standard in Sections 2.1-2.4 produce segmentations that contain no internal conflicts.[3] However, since we evaluate a gold standard by its similarity with the original annotations, it makes sense to define an approach that is capable of constructing a unified segmentation that *includes* conflicts, which we call the Union approach (UNI). UNI simply includes every segment from every annotator. The flat version ignores hierarchies (FlatUNI), and the full version includes them (FullUNI).

An example of an application of FullUNI is given in Figure 5. We see that every segment chosen by annotators 1, 2, and 3 have been included in the gold standard, creating some internal conflicts. We certainly would not expect to use this construction as a prediction of the actual gold standard, but we include it for comparison with CFU in evaluating the importance of avoiding internal conflicts with respect to our metrics.

2.7 Best Annotator

The final approach we considered chooses the "best" annotation and considers it to be the gold standard. We select the annotation with the highest inter-labeler reliability with all the other annotations to be the "best" annotation, using the pairwise κ metric. We discuss this metric and its uses in Section 3.

3 Measures of Evaluation

There are several ways of evaluating an algorithm for creating a gold standard, just as there are several ways of evaluating any segmentation algorithm. Ideally, we would like to compare to some objectively true gold standard, but it is impossible to determine if there are one or more true standards, or even if one exists. Instead, we can compare a gold standard against each annotator's individual structuring, or against that of several human annotators collectively. Also, we could compare gold standards with each other in terms of how they affect the out-

[3] MCNS avoids conflicts because any two segments that a majority of annotators agree upon will always both be included by at least one annotator, and we assume that individual annotations are always internally consistent.

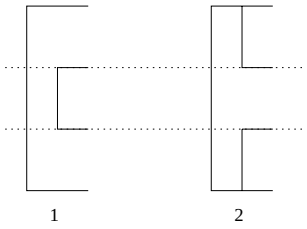

Figure 6: The κ metric would consider these two segmentations in agreement.

come of some computational task which considers discourse structure, such as anaphora resolution. This last approach is probably the best when the purpose of the gold standard is known in advance, but in this paper we consider only task-independent metrics.

For the sake of scientific validity, we did not compare a gold standard with a segmentation of our own. Instead, we chose to evaluate gold standards by averaging their similarity to the original segmentations made by human annotators. For each approach we presented earlier, we report an average similarity score over all original segmentations and the gold standard, based on several different quantitative measures of inter-reliability.

3.1 Pairwise Agreement Scores

For linear segmentation, pairwise agreement between annotators is computed by dividing the number of utterances for which both annotators agree by the total number of utterances. In contrast, a hierarchical segmentation for a sequence of utterance in a discourse is analogous to a parse tree for a sequence of words. It requires a different metric for pairwise agreement that considers the hierarchy.

Following Flammia and Zue (1995), we define a general symmetric metric P_o for observed agreement between two segments that accounts both for deletions and for insertions of segments in a hierarchy. Intuitively, we want different sub-trees that vary only in hierarchical structure but share the same boundaries to0 be considered similar. For example, in Figure 6, there is good reason to consider both annotations to be similar, even though no segment pair in either spans the same utterances.

Formally, let S_1 and S_2 be two possible segmentations. A segment $\langle i, j\rangle$ in S_1 *matches* with segmentation S_2 if there exists some segment $\langle i, *\rangle$ or $\langle *, i-1\rangle$ in S_2 and there exists some segment $\langle *, j\rangle$ or $\langle j+1, *\rangle$ in S_2. In other words, a segment in S_1 matches a segmentation S_2 if the utterances that constitute its boundaries also constitute boundaries for some segment in S_2. For example, in Figure 5, we consider that the segments H, I, and J in annotation 3 match the segments A, B, C, and D in annotation 1.

Flammia and Zue then let O_{S_1} be the number of segments in S_1 that match with segments in S_2 and let O_{S_2} be the number of segments in S_2 that match with segments in S_1. N_{S_1} and N_{S_2} are the number of segments in S_1 and S_2 respectively. Following Bakeman and Gottman (1986), they define the observed agreement to be

$$P_o = \frac{O_{S_1} + O_{S_2}}{N_{S_1} + N_{S_2}}$$

For the metric to be valid, they also take into account the probability of chance agreement between annotators. For example, if the distribution underlying the segmentation is skewed such that the structure is very sparse, most segmentations will include very few constituents, and P_o will be unnaturally deflated.

The kappa coefficient (κ) is used for correcting the observed agreement by subtracting the probability P_c that two segments in S_1 and S_2, chosen at random, happen to agree. The κ coefficient is computed as follows:

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

Carletta (1996) reports that content analysis researchers generally think of $\kappa > 0.8$ as "good reliability," with $0.67 < \kappa < 0.8$ allowing "tentative conclusions to be drawn."

All that remains is to define the chance agreement probability P_c. Let $P_b(x)$ and $P_e(x)$ be the fraction of utterances that begin or end one or more segments in segmentation x respectively. Flammia and Zue compute an upper bound on P_c as

$$P_c = \frac{N_{S_1}}{N_{S_1} + N_{S_2}} P_b(i) P_e(i) P_a(j) + \frac{N_{S_2}}{N_{S_1} + N_{S_2}} P_b(j) P_e(j) P_a(i)$$

where $P_a(x) = (P_b(x) + P_e(x) - P_b(x)P_e(x))^2$.

3.2 Precision and Recall

We use standard evaluation metrics from information retrieval to measure pairwise agreement between gold standards and annotations. We say that segment $s = \langle i, j, *\rangle$ in some segmentation *flatly agrees* with segmentation S if there exists a segment $t = \langle i, j, *\rangle$ in S, which spans exactly the same utterances as s. We say that segment $s = \langle i, j, l\rangle$ in some segmentation *fully agrees* with segmentation S if s flatly agrees with S, and the segment that fits it is also of the same depth as s; i.e., there exists a segment $t = \langle i, j, l\rangle$ in S.

We define the number of *relevant* segments in a segmentation S to be the total number of segments in S that *agree* with a gold standard for that particular discourse. For gold standard types that consider embeddedness, such as Full Consensus and Full Majority Consensus, we check for full agreement. For gold standard

types that do not, such as Flat Consensus and Conflict-Free Union, we consider flat agreement.

We define *recall* as the number of relevant segments in S divided by the total number of segments in S. We define *precision* as the number of relevant segments in S divided by the total number of segments in the gold standard. Intuitively, if a gold standard has low agreement with the original segmentation, recall will be low. If a gold standard's structure is more dense then the original segmentation, precision will be low.

3.3 Non-Crossing-Brackets

The non-crossing-bracket measure is a common performance metric used in syntactic parsing for measuring hierarchical structure similarity. A segment constituent $s = \langle i, j \rangle$ in some segmentation crosses brackets with segmentation S if s spans at least one boundary utterance in S.

For each segmentation S, we define the number of non-crossing-brackets as the number of segments in S that do not exhibit crossing brackets with the appropriate gold standard. For each segmentation, we compute a non-crossing-bracket percentage by dividing the number of non-crossing-brackets by the total number of bracket pairs.

4 Empirical Methodology

4.1 Boston Directions Corpus

For our empirical analysis of different gold standard approaches, we used the Boston Directions Corpus (BDC). The BDC corpus contains transcribed monologues by speakers who were instructed to perform a series of direction-giving tasks. The monologues were subsequently annotated by a group of subjects according to the Grosz and Sidner (1986) theory of discourse structure. This theory provides a foundation for hierarchical segmentation of discourses into constituent parts. Some of the subjects were experts in discourse theory and others were naive annotators. In our experiments here, we only consider the annotations from experts.

4.2 Experimental Design

Our experiments were run on 12 discourses in the spontaneous speech component of the BDC. The lengths of the discourses ranged from 15 to 150 intonational phrases. Each discourse was segmented by three different annotators, resulting in 36 separate annotations. For each discourse, we combined the three annotations into a gold standard according to each technique described in Section 2. We then proceeded to compute the similarity between the gold standard and each of the original annotations by using the pairwise evaluation metrics described in Section 3.

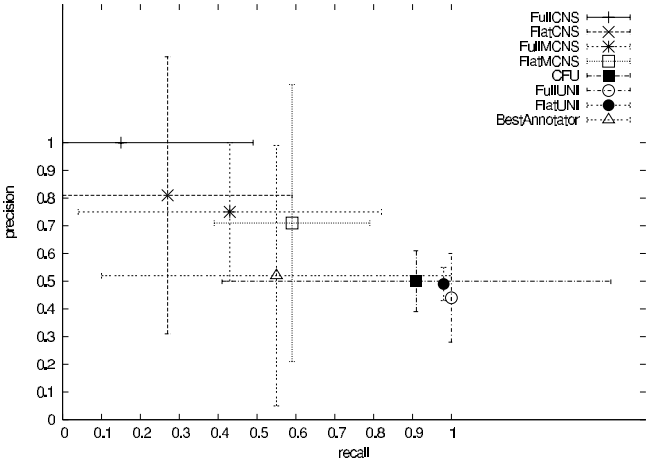

Figure 7: Results — Pairwise Agreement Scores

4.3 Results

We report results for each gold standard averaged over all 36 annotations. Table 1 presents precision/recall percentages for pairwise agreement scores, as well as κ values and non-crossing brackets (NCB) percentages. Figure 7 plots the pairwise agreement precision/recall values on a graph, with error bars indicating one standard deviation from the mean. Recall that the gold standards we are comparing are Full Consensus (FullCNS), Flat Consensus (FlatCNS), Full Majority Consensus (FullMCNS), Flat Majority Consensus (FlatMCNS), Conflict Free Union (CFU), Full Union (FullUNI), Flat Union (FlatUNI) and Best Annotator.

Our results show that CFU, FullUNI and FlatUNI all achieved high κ scores and low variance. Both Full and Flat Consensus scored the worst. This pattern was also apparent with regard to agreement between the gold standard and the annotations. Again, CFU, FullUNI and FlatUNI achieved the best recall, and FullCNS and FlatCNS scored the worst recall. It is interesting to point out that since any segmentation proposed by an evaluator will always be included in the FullUNI gold standard, its agreement recall will always be 1.

We see a change in trend with regard to precision between gold standard and the annotations. Here, FullCNS and FlatCNS achieved very high precision, while FullUNI and FlatUNI achieved low precision. CFU's precision was slightly better. With respect to the non-crossing-brackets metrics, the gold standards based on consensus (FullCNS, FlatCNS) did not clash at all with any annotation, since any segment in the gold standard is present in each of the annotations. Of the remaining methods, FlatMCNS (0.84) and FullMCNS (0.81) had the highest percentage of non-crossing-brackets, while the union based approaches, FullUNI (0.47) and FlatUNI (0.54)

	κ		Agreement Rec.		Agreement Pre.		NCB	
	ave.	sd.	ave.	sd.	ave.	sd.	ave.	sd
FullCNS	.25	.63	.15	.34	1	0	1	0
FlatCNS	.48	.42	.27	.32	.81	.60	1	0
FullMCNS	.67	.32	.43	.39	.75	.25	.81	.32
FlatMCNS	.79	.21	.59	.20	.71	.89	.84	.12
CFU	.84	.08	.91	.89	.50	.11	.78	.19
FullUNI	.84	.08	1	0	.44	.16	.53	.20
FlatUNI	.84	.09	.98	.01	.49	.06	.46	.09
BestAnnotator	.85	.15	.55	.45	.52	.47	.62	.33

Table 1: Experimental results.

had the lowest, because their gold standards are densely structured and internally include conflicts.

Looking at each gold standard separately, we do not identify a single gold standard that does well across the board. CFU, FullUNI and FlatUNI have high κ and agreement recall values, but they all have low agreement precision values. FullCNS and FlatCNS have low κ and recall values, but better agreement precision values. FullMCNS and FlatMCNS average out the best across all metrics, but they do not achieve the best performance in any of the metrics. Note that "full" type methods require agreement in hierarchy; they are held to a higher standard of evaluation than "flat" type methods.

5 Discussion

From the results, we see that generally the consensus-type approaches (CNS and MCNS) perform very well with the precision metric and the union approaches (CFU, UNI) perform well with the recall and κ metrics. Precision measures the percentage of the gold standard that was agreed upon by the annotators, and since the consensus approaches tend to include only those segments labeled by everyone, they have high precision. Specifically, FullCNS performs perfectly in precision because it contains only those segments explicitly included by everyone, while the majority consensus methods perform slightly worse because an annotator is occasionally in the minority.

Recall measures the percentage of the annotator's segments captured by the gold standard. Since the union approaches include every or almost every segment, depending on whether it is "flat" or "full," respectively, an annotator's segment is almost always included in the gold standard, yielding high recall for these methods. The difference between precision and recall highlights two different approaches: precision encourages a bottom-up approach where the most likely segments to be included in the gold standard are added from scratch; recall encourages a top-down approach where all possible segments are added and the least likely segments to be included in the gold standard are pruned. The κ metric attempts to balance these two approaches by rewarding agreements yet penalizing extra structure. Nevertheless, even the naive union methods (UNI) performs well with κ, indicating that it favors agreement far more than it punishes extra structure.

Based on these observations, we believe that there is good reason to prefer to use CFU as a gold standard over FullUNI and FlatUNI. Although they all have the same κ and similar precision/recall values, the CFU gold standard corresponds to a true segmentation — it does not exhibit internal conflicts.

However, if a conservative but accurate gold standard is desired, then the MCNS approaches are the best all-around consensus approaches to use, as they perform fairly well with κ as well as with precision and recall. These approaches construct fairly conservative gold standards, but not nearly as strict as the full consensus approaches. Hence, as seen by the high precision value, a gold standard constructed by an MCNS method will contain mostly relevant segments but will be missing the more controversial segments.

The Best Annotator approach performed very well with κ, but not as well with respect to precision and recall. Its performance was completely dominated by the MCNS approaches in all metrics, except for κ. In general, κ is at its highest when minor boundary disagreements are infrequent, because it is not sensitive to the exact type of matching boundaries. This phenomenon is shown in Figure 6. There, we see two segmentations that are clearly different but are considered the same by κ. Precision and recall, however, would not consider the second level segments in agreement.

The consistently good results of the non-crossing-brackets metric for MCNS and CFU indicate that there are few cases in which the expert BDC annotators create segments whose boundaries cross. Again, this effect is probably a result of the well-structured nature of the tasks in the BDC discourses. Since there are few cross-

ing boundaries, the κ metric performs well for the Union and Best Annotator methods since almost every boundary is represented. If annotations had exhibited more discrepancy, the non-crossing-brackets and κ metrics would probably differentiate more among these approaches.

Lastly, we note that the difference between "full" and "flat" metrics of the same type were insignificant, but with the consensus approaches, the "flat" approaches performed slightly better than their "full" counterparts, most likely because the "full" approaches were too conservative in demanding level agreement. Thus, if we care not to have conflicts in our gold standard, the "full" approaches should be used to find the gold standard, as they produce more structured segmentations. In addition, a gold standard with labeled embeddedness might be necessary for post-segmentation processing, such as anaphora resolution. However, if the gold standard is being used for purely evaluative reasons, the "flat" approaches should be used as they perform slightly better.

6 Future Work

One problem with the measures of evaluation that we have explored in this paper is that they tell us how similar a gold standard is to the original annotations but say nothing about how effective the gold standard would be when used for further discourse processing. One suggestion for future studies would be to evaluate the gold standards with respect to possible post-segmentation tasks, such as anaphora resolution or summarization. Such an approach would be a better measure of the objective goodness of the gold standard and could also be a way to monitor the skills of the annotators. Specific metrics might also be more relevant for a specific discourse task. For instance, perhaps non-crossing-brackets is a more useful metric to consider when segmenting as a preprocessing step for anaphora resolution.

It would also be interesting to further explore the Conflict-Free Union approach, as it performed well but suffered from including extra structure. The top-down processing could be enhanced by removing those segments which are deemed the least probable to be in the gold standard, perhaps based on some features, such as depth. For example, perhaps a segment that is at a deep level and is only in a few annotations could get removed, while larger segments would remain regardless, or vice versa. With a few good features, it seems quite possible to increase the precision of CFU. A similar approach could be taken to add new segments to those picked in the Majority Consensus approach.

Finally, it is worth exploring whether it is a good idea to have multiple annotations for a given corpus in the first place. Some corpora, such as the Penn Treebank, require its annotators to meet whenever there is a conflict so that the conflict can be resolved before the corpus is publicly released. Penn has now begun a Discourse Treebank as well (Creswell et al., 2003). Wiebe et al. (1999) use statistical methods to automatically correct the biases in annotations of speaker subjectivity. The corrections are then used as a basis for further conflict resolution. Carlson et al. (2001) also used conflict resolution when creating their discourse-tagged corpus. One interesting area of research would be to compare how annotators choose to resolve their conflicts compared to the different automatic approaches of finding a gold standard. It is possible that the compromises made by the annotators cannot be captured by any computational method, in which case it may be worth having all conflicts resolved manually.

Acknowledgments

We would like to thank Jill Nickerson for comments on an earlier draft of this paper. This work is supported in part by the National Science Foundation under Grant No. IIS-0222892, the GK-12 Fellowship (NSF 04-533), and NSF Career Award IIS-0091815.

References

R. Bakeman and J.M. Gottman. 1986. *Observing interactions: an introduction to sequential analysis*. Cambridge University Press.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2).

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proc. of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Denmark, September.

C. Creswell, K. Forbes, E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2003. Penn discourse treebank: Building a large scale annotated corpus encoding dltag-based discourse structure and discourse relations. *Manuscript [fix this]*.

G. Flammia and V. Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Proc. Euroospeech-95*, volume 3, pages 1965–1968, Madrid, Spain.

B.J. Grosz and J. Hirschberg. 1992. Some intonational characteristics of disourse structure. In *Proc. of ICSLP-92*, volume 1.

B.J. Grosz and C.L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

M. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33–64.

J. Hirschberg and C. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of ACL-1996*, Santa Cruz.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, November.

R.J. Passonneau and D.J. Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Meeting of the Association for Computational Linguistics*, pages 148–155.

J.A. Rotondo. 1984. Clustering analysis of subject partitions of text. *Discourse Processes*, 7:69–88.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland, June.